图书情报工作 ChinaXiv合作期刊
第 63 卷 第 11 期　2019 年 6 月

# 发文趋势与引文趋势融合的学科研究主题优先级排序*
## ——以我国情报学学科主题为例

李秀霞[1]　程结晶[2]　韩霞[1]
[1] 曲阜师范大学传媒学院　日照 276826　[2] 扬州大学社会发展学院　扬州 225008

**摘要**:［目的/意义］主题排序不仅是信息检索、信息组织研究的基础性问题,也是图书馆学科服务的重要工作,对学科领域研究主题进行有效排序能够帮助科研人员和科研管理部门有效把握学科领域的研究态势,准确定位科研方向,快速做出科研决策。［方法/过程］基于趋势分析提出一种学科研究主题优先级排序算法。首先,在主题提取的基础上,根据发文趋势和引文趋势将每个研究主题按研究等级分为贫乏主题、热点主题、冷点主题、过热主题 4 个子类。然后,分别对各子类下的主题词进行优先级排序。［结果/结论］在情报学领域的实验表明:本文提出的优先级排序算法能够全方位、细粒度、深层次地展示学科领域研究主题的发展等级,该方法可为从时间维度实现动态情报分析提供新的视角。

**关键词**:发文趋势　引文趋势　研究主题　优先级排序
**分类号**:G250.2
**DOI**:10.13266/j.issn.0252-3116.2019.11.010



## 1　引言

随着学科研究的深入和跨学科研究的拓展,学术文献呈爆炸式增长态势,研究主题不断演化更新。面对主题多样的海量文献资源,如何迅速、准确地掌握学科研究主题的发展等级,确定科研选题方向,成为科学研究者面临的巨大挑战[1]。目前,不少学者以文献篇章为基本单元,通过文献排序提供信息服务、指导科研工作[2-3],文献排序虽然能够给出学科领域有价值的文献、权威作者等信息,但面对学科研究的不断深入与拓展,这种信息服务远远不够。由此,本研究基于趋势分析提出一种学科主题优先级排序方法,对文献内容进行深入挖掘,为科研人员提供层次更深、细粒度更高的信息服务[4]。

本研究的主要目标是:①给出相对引文量、引文趋势、发文趋势的定义,为从时间维度动态分析学科研究主题的发展演化提供理论基础;②在趋势分析的基础上,给出学科研究主题优先级排序方法,为发现学科领域研究主题的精细发展态势提供理论支持;③利用提出的研究主题排序算法对我国情报学研究主题进行优先级排序,为本学科科研管理人员制定科研规划、研究人员进行科研选题等提供有效、可靠的决策参考。

## 2　相关研究

### 2.1　基于文献计量学的主题识别

这类研究包括词频分析[5]、共词分析[6]、共词聚类分析[7]等,上述方法实质上都是以高频关键词为基础,识别文献簇的研究主题,进而发现学科领域的研究热点。关键词是文献研究主题、研究内容、研究方法的高度概括与凝练,反映文献研究的逻辑关系或创新突破点;高频关键词则代表着一个学科领域的热点主题和前沿方向,频次越高的关键词得到的研究关注度就越高,通常构成研究热点[8]。因此可通过统计和分析关键词在文献中出现的频次高低来确定学科领域的研究热点和发展趋势。基于文献计量学的主题识别因技术方法具有较强的通用型、分析工具简单易用而被广泛

* 本文系国家社会科学基金项目"文献内容分析与引文分析融合的知识挖掘与发现研究"(项目编号:16BTQ074)研究成果之一。
作者简介:李秀霞(ORCID:0000-0002-3492-4768),教授,硕士生导师,E-mail:zyshao@126.com;程结晶(ORCID:0000-0003-0158-7854),教授,博士生导师;韩霞(ORCID:0000-0003-2000-5776),硕士研究生。
收稿日期:2018-08-12　修回日期:2018-12-19　本文起止页码:88-95　本文责任编辑:王传清

应用于学科领域研究热点识别和研究结构分析中。不足在于关键词的选取主观性强、关键词之间缺乏语义关系、会遗漏频次较低且代表新兴研究主题的关键词等，致使这类方法在揭示领域知识结构时效果不够理想[9]。

### 2.2 基于机器学习的主题挖掘

主题挖掘源于 G. Salton 等[10]于 1975 年提出的向量空间模型（Vector Space Model, VSM），VSM 将文本表达成几何空间中的向量，为计算文本之间的相似度、确定关键词与文本的关系提供了便利。1990 年 S. C. Deerwester 等[11]提出的潜在语义分析（Latent Semantic Analysis, LSA）模型首次成功地将“语义”引入文本主题挖掘。1999 年，T. Hofmann[12]运用期望最大化算法提出了基于概率统计的 PLSA（Probabilistic Latent Semantic Analysis）模型，将机器学习纳入文本主题提取。2003 年，D. M. Blei 等[13]在 PLSA 的基础上，把先验概率引入隐含语义分析中，提出潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）模型。LDA 模型假设词是由一个个主题混合产生，同时每个主题是在固定词表上的一个多项式分布，这些主题被集合中的所有文档所共享，每个文档有一个特定的主题比例，从 Dirichlet 分布中抽样产生。在实际应用中，只要确定了文档集、明确隐含主题的数量，即可实现主题提取。目前，LDA 模型已成为广泛使用的一种主题挖掘模型，并衍生出一系列的主题挖掘方法[14－17]。相比基于文献计量分析的主题识别，基于机器学习的主题挖掘不仅在主题识别中能够挖掘出更多、更全面的内容，内容描述更具体、明确[18－19]，而且主题内关键词的语义联系更为紧密，对语义关系模糊、逻辑结构粗糙的文献，挖掘正确主题的效果更理想[20]。

### 2.3 基于机器学习的主题排序

基于机器学习的主题排序是在主题识别的基础上，通过构建排序模型，计算主题的相关度对其进行排序，目前已被应用于文档检索、协同过滤、专家搜索、情感分析、产品评级等。也有对新闻主题、社交媒体主题的排序研究，如 C. H. Wang 等[21]采用媒体聚焦和用户注意力的方式对主题排序；姜晓伟等[22]根据微博话题的影响力、突发性和相关性，结合 LDA 主题模型实现了对微博重要话题的发现与排序，刘培玉等[23]对微博文本和主题词的热度进行联合排序，用于微博热点主题词的抽取以及热点话题的发现。对学科主题排序的研究相对较少，代表性的研究有：W. Cui 等[24]借助信息检索与数据挖掘中常用的 TF-IDF 加权技术，提出利用 TF-IDF 对主题词排序，该方法对于词语比较少的情况效果较好，但面对高维数据集则显得力不从心。之后，出现了一些主题排序模型，解决了高维数据主题排序的问题，如肖智博[25]提出一种基于关联关系的主题排序模型算法，该算法利用主题之间的各种关联关系，依照主题的重要性程度排序。后来，肖智博与他的学生[26]研发了一种基于排序主题模型的论文推荐系统。另有学者借鉴网页排序的 PageRank 算法实现对科技主题的排序，如蒋卓人等[27]借助 PageRank 算法对中英文科技主题的重要性进行了度量和排序，由于 PageRank 算法是基于链接分析的，不能很好的基于主题查询，因此计算结果往往会偏离实际的查询主题。

对学科领域研究主题进行有效排序能够帮助科研人员和科研管理部门有效把握学科领域的研究态势、准确定位科研方向、快速做出科研决策，意义重大，应用广泛。但目前基于算法模型对学科研究主题的排序研究主要是在文本挖掘的基础上，根据主题词出现的频次或主题词间的关联性实现主题排序，尚未发现有考虑用户需求因素的相关研究。为此，本文在前人研究的基础上，将文献计量和主题挖掘两种方法相结合，从读者和研究人员两个视角、通过发文趋势和引文趋势两个维度实现对学科研究主题的合理排序。

## 3 研究步骤与研究方法

### 3.1 主题提取与主题数目确定

主题提取即提取学科领域学术文献的研究主题。一般而言，学术文献的标题能够提供文献的核心问题，如研究内容、研究方法、研究目标等；关键词则是对文献核心内容的高度概括。如前所述，LDA 模型具有良好的文本潜在主题挖掘能力，能够识别大规模文档集或语料库中潜藏的主题信息[22]，目前已被应用于主题抽取、热点挖掘、文本分类、用户推荐等领域。因此，本文确定采用 LDA 模型，从文献标题和关键词中提取学科领域学术文献的研究主题。

在学科文献主题提取中，主题数目的确定至关重要，主题数目过少不能涵盖学科领域的研究全貌，过多则会出现重复分析的现象；而根据作者或专家建议确定主题数目又带有主观性的弊端。因此，本文利用所有主题之间的平均相似度来度量主题结构的稳定性，平均余弦值在 1 和 0 之间，主题之间的平均相似度越小，对应的主题结构越优[28]。

### 3.2 主题词引文等级确定

为便于描述，针对某一学科做如下假设：设有 $M$

个主题,其中任一主题 $m$ 含 $k$ 个主题词,某年对应某一主题词共有 $N$ 篇论文。

首先,统计某年某个主题词的相对引文量 $R_{Cj}$;然后,计算学科领域某年所有主题词上的相对引文量 $T_C$;再根据 $R_{Cj}$、$T_C$ 的值计算样本标准差 $d$;最后,确定某年第 $j(j=1,2,\cdots,k)$ 个主题词的引文等级 $q_{jm}$。其中,

$$R_{Cj}=\frac{\sum_{i=1}^{N}C_i+1}{N+1} \qquad \text{公式(1)}$$

$R_{Cj}$表示某年第 $j(j=1,2,\cdots,k)$ 个主题词对应文献的所有引文量与同年该主题词对应的所有文献量之比。式中 $C_i$ 代表第 $i(i=1,2,\cdots,N)$ 篇文献的引文量。为避免发文量或引文量为 0 时无法计算,分子分母同时加 1。

$$T_C=\frac{\sum_{j=1}^{k}\sum_{i=1}^{N}C_{ij}}{\sum_{j=1}^{k}\sum_{i=1}^{N}P_{ij}} \qquad \text{公式(2)}$$

$T_C$ 表示某年所有主题词对应文献的引文量与该年主题 $m$ 下所有主题词对应的文献量之比。式中 $P_{ij}$ 表示第 $j$ 个主题词的第 $i$ 篇文献,$C_{ij}$表示第 $j$ 个主题词的第 $i$ 篇文献的引文量。

$$d=\sqrt{\frac{\sum_{j=1}^{k}(R_{Cj}-T_C)^2}{K}} \qquad \text{公式(3)}$$

$d$ 表示某年 $k$ 个主题词的样本标准差。

$$q_{in}=\left(\frac{R_{Cj}-T_C}{d}\right) \qquad \text{公式(4)}$$

$q_{jn}$表示某年第 $j$ 个主题词的引文等级,$n$ 代表时间段($n=1,2,\cdots$),各主题词每年都有一个引文等级。

## 3.3 主题词引文趋势、发文趋势

根据主题词的引文等级,构建各主题词 $n$ 个时间段的引文等级向量,即 $Q_{jn}=(q_{j1},q_{j2},\cdots,q_{jn}$;各主题词均对应相同的时间向量,即 $Y=(y_1,y_2,\cdots,y_n)$。

对各主题词引文等级向量 $Q_{jn}$与时间向量 $Y$ 进行 Spearman 相关分析,得到各主题词相对时间的 Spearman 相关系数 $Q_{Rj}$。Spearman 相关系数表明了引文等级向量 $Q_{jn}$和引文时间 $Y$ 的相关方向。如果 $Y$ 增加,$Q_{jn}$趋向于增加,则 $Q_{Rj}$为正;如果 $Y$ 增加,$Q_{jn}$趋向于减少,则 $Q_{Rj}$为负;$Q_{Rj}$为零则表明当 $Y$ 增加时 $Q_{jn}$没有任何趋向性。$Q_{Rj}$的大小反映了读者对各主题词的需求增长或减少的趋势,以此记为各主题词的引文趋势。

统计每个主题词不同时间段的发文量,以“时间”为行、以“主题词”为列构建“发文量 - 时间”矩阵,各主题词每个时间段的发文量用向量 $L_{jn}$表示,$L_{jn}=(l_{j1},l_{j2},\cdots,l_{jn})$。

对各主题词的发文量向量 $L_{jn}$与时间向量 $Y$ 进行 Spearman 相关分析,得到各主题词与发文时间的 Spearman 相关系数 $L_{Rj}$。Spearman 相关系数反映了 $L_{jn}$和发文时间 $Y$ 的相关方向。$L_{Rj}$为正,说明当 $Y$ 增加时,$L_{jn}$有增加的趋势;$L_{Rj}$为负,说明 $Y$ 增加时,$L_{jn}$有减少的趋势;$L_{Rj}$为零,表明 $L_{jn}$没有任何变化趋向性。相关系数 $L_{Rj}$的大小反映了研究人员对各主题词的研究递增或递减趋势,以此记为各主题词的发文趋势。

## 3.4 主题词优先级排序

3.4.1 主题词优先级划分 根据相对引文量、发文趋势、引文趋势的定义,计算学科领域研究主题对应主题词的发文趋势 $L_{Rj}$、引文趋势 $Q_{Rj}$。根据 $L_{Rj}$与 $Q_{Rj}$的不同取值,将研究主题细分为 4 类子主题,各类子主题分别代表着不同的研究等级。分类标准如下:①当一个主题词对应的发文趋势降低、引文趋势增加时,说明需求量大于供给量,相关研究处于贫乏状态,急需增加研究量。因此,界定这一类主题词属于研究贫乏的主题词,对这种主题词的相关研究急需给予引导和支持,研究级别最高。②当一个主题词对应的发文趋势和引文趋势均递增时,说明需求量相对快速增加时,供应量也在高速递增,属于学科领域的热点主题词,对该类主题词的研究能够满足需求,因此,研究级别应低于贫乏区的主题词。③当一个主题词对应的发文趋势和引文趋势均递减时,说明对该研究主题的需求量和供给量都较低,属于冷点研究主题词。相对需求较低的主题词,不需要给予过多支持,故研究级别又低于热点主题词。④当一个主题词对应的发文趋势增加,引文趋势降低时,供给量大于需求量,说明对该主题词的研究增幅相对较快,呈现研究过热的势头。因此,对该类主题词的研究应该进行适当控制,故研究级别最低。具体表示为:

$$\begin{cases} L_{Rj}<0,Q_{Rj}>0,(\text{①类,贫乏主题}) \\ L_{Rj}>0,Q_{Rj}>0,(\text{②类,热点主题}) \\ L_{Rj}<0,Q_{Rj}<0,(\text{③类,冷点主题}) \\ L_{Rj}>0,Q_{Rj}<0,(\text{④类,过热主题}) \end{cases} \qquad \text{公式(5)}$$

3.4.2 子主题排序 通过对发文趋势 $L_{Rj}$和引文趋势 $Q_{Rj}$的运算实现各类子主题下主题词的优先级排序。排序依据自定义运算关系 $r_j=L_{RJ}\odot Q_{RJ}$进行,“ ⊙”是一种自定义运算符[28],应用时需根据数据的不同分布特点自行定义 $L_{Rj}$与 $P_{Rj}$之间的运算关系。

## 4 实验与效果评价

### 4.1 数据来源

本文数据选自中文社会科学引文索引(CSSCI)收录的来源期刊。CSSCI 期刊是目前我国社会科学各学科领域具有较高学术水平的期刊，刊载在这些期刊上的文献基本涵盖了各学科领域的研究主题。其中，“图书馆、情报与文献学”期刊有 20 种，20 种期刊中情报学期刊(含图书馆学与情报学两栖期刊)有 10 种，分别是《情报学报》《图书情报工作》《情报杂志》《图书情报知识》《情报资料工作》《数据分析与知识发现》《情报理论与实践》《情报科学》《图书与情报》《现代情报》。为同时获取上述期刊的提名和关键词以备后面的主题提取，笔者以中国知网（CNKI）为来源数据库对上述 10 种情报学期刊文献进行全面检索。检索时间范围为 2013 年 6 月至 2018 年 5 月，共检索到 13 559 篇文献，剔除其中的会议通知、下期目录、征稿通知等无关数据，得到有效文献 12 377 篇。下载这些文献的题名、关键词等信息作为实验数据。

### 4.2 我国情报学研究主题数目确定及主题提取

利用 LDA 模型进行主题提取之前，需对数据进行预处理。首先，笔者使用中国科学院计算技术研究所汉语分词系统 NLPIR（又名 ICTCLAS）对样本数据进行分词处理；然后，使用词性过滤和停用词过滤方法对与建模无关的词语进行过滤，得到实验所需的文本语料库。

选取不同的主题数目，计算主题间的平均相似度，发现当主题数为 10 时，主题间的平均相似度最小，主题结构最稳定，具体如图 1 所示：

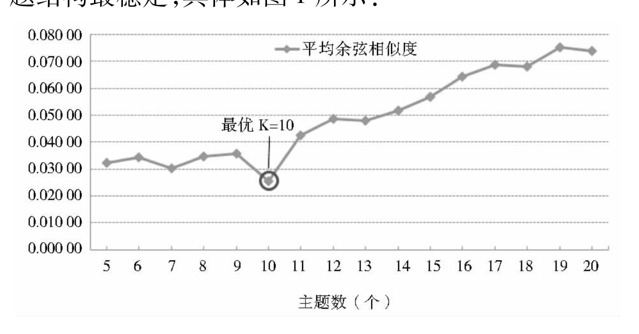



图 1 我国情报学主题数目与主题间平均余弦相似度关系曲线

利用开源包 JGibbLDA[27] 对数据进行 LDA 主题建模，设置主题数为 10，α 和 seita 分别设置为 0.1 和 0.02，提取主题。根据 LDA 模型生成的各研究主题列表中的主题词，并依据笔者对学科领域研究主题的把握，由人工确定 10 个研究主题的标签，主题标签见表 1。

### 4.3 10 个研究主题下各子类主题优先级划分

为了与上文数据保持一致，选用 CNKI 数据库文献分类目录中“信息科技”类目下的“情报学、情报工作”。通过“更多”选择“指数”。首先，对上述各类主题下的主题词进行简单的合并处理，如“比较分析”与“对比分析”，“网络舆情”“网络舆论”与“网络舆论传播”，“开放政府数据”与“政府开放数据”，“满意度”与“用户满意度”，“可视化”与“可视化分析”，“评价指标体系”与“指标体系”等的合并。然后，将处理后的主题词逐一输入，依次检索并记录各主题词每年对应的“学术关注度”(即发文量)和“学术传播度”(即引文量)。结合主题提取时样本数据检索时间(2013 年 6 月至 2018 年 5 月)，考虑引文相对发文的时滞性，设定发文时间为 2013 年 6 月至 2017 年 5 月，引文时间设定为 2014 年 6 月至 2018 年 5 月。根据 3.2 节、3.3 节给出的相对引文量、发文趋势、引文趋势的定义，计算各类主题词的发文趋势 $L_{Rj}$、引文趋势 $Q_{Rj}$。根据 $L_{Rj}$ 的不同取值，按照 3.4 节给出的子主题分类方法将各研究主题下的主题词分为贫乏主题、热点主题、冷点主题、过热主题等 4 个子类(部分研究主题分为 3 个子类)，分类结果见表 1。

### 4.4 10 个研究主题下各子类主题的主题词排序

利用上文给出的研究主题优先级自定义排序算法，分别对不同研究主题下各子类的主题词进行优先级排序。分析各子类主题词的发文趋势 $L_{Rj}$、引文趋势 $Q_{Rj}$ 特点：有的主题词对应的 $Q_{Rj}$ 与 $L_{Rj}$ 呈负相关，有的主题词对应的 $Q_{Rj}$ 与 $L_{Rj}$ 呈正相关；另外，引文趋势 $Q_{RJ}$ 多聚集在(－1，－0.7)以及(0.7，1)范围，最大值与最小值差距较大。同时，为使 4 个子类的 $r_j$ 值的大小排序与其优先级高低变化保持一致，并确保所有主题词的 $r_j$ 值为正，经反复实验，本文设计以下的优先级排序算法：

$$\begin{cases} r_j = 13Q_{Rj} - L_{Rj} + 14, (Q_{Rj}\text{与 }L_{Rj}\text{变化趋势相反时}) \\ r_j = 2Q_{Rj} - L_{Rj} + 14, (Q_{Rj}\text{与 }L_{Rj}\text{变化趋势相同时}) \end{cases}$$

公式(6)

按照上述自定义算法，计算各子类主题下每个主题词的 $r_j$ 值，结果见表 1。

### 4.5 效果评估

由于目前没有公认的关于学科主题排序的评估方法，更未发现对我国情报学研究主题的排序研究，本文选取排序结果合理性分析、对比实验分析两种方法来评估排序效果。

**表 1 基于趋势分析的情报学研究主题优先级排序**

| 子类 | topic0th 信息管理 | rj 值 | 子类 | topic1th 信息服务 | rj 值 | 子类 | Topic 2th：网络舆情 | rj 值 | 子类 | Topic 3th：网络资源 | rj 值 | 子类 | Topic 4th：专利信息 | rj 值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ①类 | 科技查新 | 25.051 9 | ①类 | 电子资源 | 25.635 3 | ①类 | 网络环境 | 26.195 6 | ①类 | 领域本体 | 26.425 6 | ①类 | 专利分析 | 26.252 5 |
| | 行为研究 | 24.590 3 | | 用户需求 | 19.426 4 | | 图书馆联盟 | 26.160 6 | | 个性化 | 26.283 7 | | 信息共享 | 26.207 1 |
| | 理论基础 | 24.345 9 | | 利用率 | 16.324 9 | | 微博舆情 | 25.252 4 | | 信息资源 | 25.359 8 | | 专利申请 | 25.946 8 |
| | 比较分析 | 20.666 4 | ②类 | MOOC | 15.113 3 | | 网络平台 | 24.691 6 | | 知识组织 | 23.865 7 | | 理论基础 | 24.084 1 |
| | 信息管理 | 15.845 2 | | 服务模式 | 14.294 6 | | 信息需求 | 21.951 | | 科技文献 | 21.808 2 | | 专利权人 | 22.891 4 |
| | 学科馆员 | 15.448 4 | | 创客空间 | 14.273 9 | ②类 | 网络舆情 | 15.145 2 | | 智能化 | 20.057 4 | | 知识共享 | 21.42 |
| ②类 | 内容分析法 | 14.570 2 | | 全民阅读 | 13.996 6 | | 传播规律 | 15.763 4 | | 网络信息 | 16.320 4 | | 专利文献 | 19.707 9 |
| ③类 | 信息资源共享 | 13.345 | ③类 | 文献资源 | 12.558 | | 系统动力学 | 15.26 | | 网络资源 | 15.609 | | 知识产权 | 19.686 4 |
| | 信息行为 | 13.070 8 | ④类 | 服务能力 | 11.701 5 | | 舆情传播 | 14.725 7 | | 网络社区 | 14.365 8 | | 专利信息 | 19.108 8 |
| ④类 | 政府信息公开 | 11.656 9 | | 阅读推广 | 11.701 5 | | 舆情事件 | 14.547 5 | ②类 | 文本挖掘 | 14.977 4 | | 专利数据 | 17.183 7 |
| | 学科服务 | 11.271 4 | | 服务体系 | 7.352 98 | | 大数据 | 14.326 6 | | 模型构建 | 14.128 4 | | 隐性知识 | 16.816 6 |
| | 信息公开 | 9.951 7 | | 博物馆 | 5.102 51 | | 大数据技术 | 14.252 3 | | 资源整合 | 14.089 9 | ②类 | 核心专利 | 15.191 1 |
| | 案例分析 | 9.463 48 | | 服务内容 | 5.095 33 | | 复杂网络 | 14.067 2 | ③类 | 维基百科 | 14.085 5 | | 知识网络 | 14.588 9 |
| | 发展历程 | 6.766 79 | | 微信公众平台 | 4.095 94 | | 互联网 | 14.054 | | 知识单元 | 13.927 8 | | 协同创新 | 14.386 9 |
| | 信息素养 | 5.874 6 | | PDA | 3.998 71 | | 网络舆情传播 | 13.994 9 | | 信息检索 | 13.716 6 | | 科技创新 | 13.709 2 |
| | 理论研究 | 4.683 19 | | 移动服务 | 1.564 7 | | 新媒体环境 | 13.897 8 | | 数字资源 | 13.099 9 | ④类 | 知识交流 | 13.210 1 |
| | 电子政务 | 1.861 04 | | | | ④类 | 网络舆论 | 12.666 9 | ④类 | 信息生态链 | 13.053 5 | | 网络结构 | 13.044 8 |
| | | | | | | | 新媒体 | 11.602 2 | | 知识协同 | 13.001 6 | | 创新能力 | 11.914 8 |
| | | | | | | | 知识扩散 | 10.04 8 | | 数字化 | 8.093 66 | | 合作关系 | 11.043 4 |
| | | | | | | | 数据分析 | 6.766 22 | | 主题词 | 4.365 06 | | 识别方法 | 8.123 55 |
| | | | | | | | 数据管理 | 6.067 65 | | 情感分析 | 3.393 22 | | 竞争力 | 5.933 55 |
| | | | | | | | 传播模式 | 2.766 24 | | 语料库 | 3.227 41 | | 关联性 | 5.090 9 |
| | | | | | | | 博弈模型 | 2.210 49 | | 生命周期 | 2.990 7 | | 科技报告 | 4.378 99 |
| | | | | | | | 图书馆员 | 1.952 06 | | 网络化 | 2.789 68 | | 知识创新 | 4.351 29 |
| | | | | | | | 情报服务 | 1.361 73 | | 结构化 | 2.148 04 | | 技术创新 | 3.173 29 |
| | | | | | | | 大数据环境 | 1.128 12 | | 知识发现 | 1.979 92 | | 社会网络 | 2.409 73 |
| | | | | | | | 情报分析 | 0.981 | | LDA | 1.879 57 | | 虚拟社区 | 1.428 67 |
| | | | | | | | 互联网 + | 0.396 19 | | 机器学习 | 0.223 62 | | 网络分析 | 0.304 56 |
| 子类 | Topic 5th：开放数据 | rj 值 | 子类 | Topic 6th：信息用户 | rj 值 | 子类 | Topic 7th：情报评价 | rj 值 | 子类 | Topic 8th：文献计量 | rj 值 | 子类 | Topic 9th：信息生态 | rj 值 |
| ①类 | 平台建设 | 25.226 | ①类 | 政务微博 | 26.909 | ①类 | 学术期刊 | 26.148 8 | ①类 | 影响因子 | 26.686 2 | ①类 | 信息生态 | 26.621 3 |
| | 信息服务 | 23.28 | | 微博用户 | 25.875 | | 开放存取 | 26.126 | | 核心期刊 | 25.895 7 | | 搜索引擎 | 25.744 5 |
| | 知识管理 | 20.982 | | 信息搜寻 | 25.391 6 | | h 指数 | 25.342 3 | | 统计分析 | 21.557 6 | | 知识服务 | 18.104 8 |
| | 用户体验 | 19.504 6 | | 用户满意度 | 21.019 2 | | 学术论文 | 24.581 6 | ②类 | 期刊论文 | 15.127 7 | ②类 | 评价模型 | 14.489 6 |
| ②类 | 推荐系统 | 15.251 3 | | 网络用户 | 20.517 2 | | 图书馆学 | 23.127 6 | | 数据库 | 14.935 2 | | 服务创新 | 14.474 8 |
| | 云计算 | 15.069 4 | | 信息传播 | 18.710 8 | | 文献调研 | 21.27 | | 研究领域 | 13.910 2 | | 信息生态链 | 14.441 6 |
| | 移动图书馆 | 14.930 8 | | 用户信息 | 18.401 8 | | 信息分析 | 17.515 1 | | 被引频次 | 13.832 4 | | 知识库 | 14.290 9 |
| | 信息技术 | 14.621 | | 信息获取 | 16.473 7 | ②类 | 关联数据 | 15.158 5 | | 数据源 | 13.690 1 | | 指标体系 | 14.267 3 |
| | 对比分析 | 14.300 2 | ②类 | 社会化媒体 | 15.576 9 | | 科研人员 | 14.898 9 | | 共词分析 | 13.084 6 | | 用户参与 | 13.854 9 |
| | 机构知识库 | 14.227 4 | | 在线评论 | 15.015 | | 国家安全 | 14.562 | ④类 | 社会网络分析 | 9.702 49 | ③类 | 竞争情报 | 12.990 8 |
| | 政府数据开放 | 14.082 1 | | 用户行为 | 14.739 3 | | 科学数据 | 14.445 8 | | 文献计量 | 8.225 22 | | 信息生态系统 | 12.967 2 |
| | 服务平台 | 13.721 1 | | 信息消费 | 14.534 2 | | 科研数据 | 14.058 2 | | 跨学科 | 7.764 31 | ④类 | 层次分析法 | 11.740 7 |
| | 管理模式 | 13.680 7 | | 技术接受模型 | 14.474 | | 元数据 | 14.035 1 | | 研究主题 | 6.948 23 | | 供应链 | 11.503 1 |
| | 公共服务 | 13.444 2 | | 扎根理论 | 14.25 | | 科研数据管理 | 13.587 8 | | 相关文献 | 5.914 08 | | 应急决策 | 11.293 6 |
| | 文献分析 | 13.164 8 | | APP | 14.201 9 | ④类 | Altmetrics | 11.486 | | 聚类分析 | 4.941 21 | | 人工智能 | 8.263 25 |
| ③类 | 个人信息 | 12.746 | | 意见领袖 | 13.631 8 | | 智库建设 | 10.640 2 | | 文献计量学 | 2.200 16 | | 信息组织 | 8.172 58 |
| ④类 | 开放数据 | 12.731 6 | | 社交网络 | 13.521 1 | | 学术交流 | 9.846 26 | | 研究热点 | 1.285 82 | | 生态学 | 6.999 44 |
| | 智慧城市 | 11.156 | | 移动阅读 | 13.437 9 | | 学科交叉 | 8.844 16 | | 可视化分析 | 0.899 41 | | 应急管理 | 6.855 9 |
| | 馆藏资源 | 10.090 2 | ③类 | 知识转移 | 12.872 2 | | 开放获取 | 7.924 02 | | 知识图谱 | 0.291 | | 知识元 | 6.179 7 |
| | 电子商务 | 9.958 53 | ④类 | 移动互联网 | 12.333 9 | | 学术影响力 | 7.678 34 | | | | | UGC | 4.962 62 |
| | 可用性 | 9.567 53 | | 新浪微博 | 12.164 5 | | 科研机构 | 5.811 61 | | | | | 服务质量 | 4.511 01 |
| | 资源聚合 | 7.955 24 | | 结构方程模型 | 10.814 3 | | 评价方法 | 5.729 89 | | | | | 突发事件 | 2.754 21 |
| | 信息安全 | 7.226 85 | | 关键因素 | 10.089 9 | | 评价指标 | 3.377 65 | | | | | 指标权重 | 1.749 24 |
| | 信息系统 | 4.880 94 | | 社交媒体 | 7.663 47 | | 情报研究 | 2.552 44 | | | | | 信息环境 | 0.585 51 |
| | 数字图书馆 | 4.620 84 | | 消费者 | 6.475 9 | | 评价体系 | 1.962 61 | | | | | | |
| | 知识资源 | 2.939 5 | | 理论模型 | 5.481 76 | | | | | | | | | |
| | 运行机制 | 2.554 42 | | 概念模型 | 3.889 16 | | | | | | | | | |

4.5.1 排序结果合理性分析　下面仅以“Topic5th：开放数据”为例，分析排序结果的合理性。

该类主题下的第①子类含 4 个主题词，代表着用户对信息服务的体验、知识管理平台的建设两个子方向。目前，信息服务、知识管理的相关研究理论已近完善，缺乏创新的发展途径和理念，因此，发文量呈逐年递减趋势。但在崇尚虚拟现实环境、倡导“以人为本”理念的时代，人们越来越关注用户体验，对图书馆服务的需求由信息服务逐渐转向知识服务。可见，将该子类主题设置为最高研究级别是合理的，可以指导相关机构和部门采取一定的措施加大对这两个子方向研究的扶持力度，以满足人们对该子类研究主题日益剧增的需求。

第②子类包含的主题词主要研究的是推荐系统、云计算、移动服务、政府开放数据等问题。在互联网技术、移动技术、物联网技术迅速发展和广泛普及的信息

李秀霞，程结晶，韩霞. 发文趋势与引文趋势融合的学科研究主题优先级排序——以我国情报学学科主题为例[J]. 图书情报工作,2019,63(11):88－95.

环境下,图书情报学界高度重视数据开放及其应用平台的建设,注重移动终端在信息服务中的地位,这与上述主题词对应的发文量和引文量都逐年递增不谋而合。但如果一个学科领域过多地关注某个研究方向,则不利于学科的均衡发展。因此,设置此子类研究主题的研究级别低于第①子类,以提醒研究人员,对此类研究主题应头脑冷静,谨防出现研究过热现象。

第③子类仅有一个主题词,即个人信息,其发文量、引文量都逐年递减。虽然在网络信息时代,人们更加关注网络信息安全和用户隐私保护,但由于图书情报学不像其他学科(如计算机学科)涉及更多的隐私安全问题,因此,该子类主题成为图书情报学的研究冷点,研究级别不高自在情理之中。

第④子类的主题词代表的研究方向主要是数据化资源。资源管理与建设一直是图书情报学研究的热点,也是该学科领域最擅长的研究方向。在数字化时代,资源的数字化研究(如信息资源的数字化、资源组织与管理的数字化、资源利用的数字化等)一时间成为图书情报学领域炙手可热的研究方向,属于研究过热的主题。对此子类主题,相关机构部门(如项目审批机构、图书情报学学术期刊)应采取相应措施适当控制这部分主题的研究量,故将该子类主题的研究等级设置为最低等级。

4.5.2 对比试验分析 为方便对比,本文利用同一数据集进行基于共词聚类分析的主题词排序,即通过共词分析、共词聚类、社会网络分析等过程将主题词排序。

具体过程是:将上述 10 种期刊的题录信息(包括关键词)导入 bicomb 中,提取每一篇文献的关键词,通过合并、删减等规范化处理后,共得到 27 057 个关键词。选取出现频次大于等于 20 次的 244 个高频关键词生成共词矩阵,通过相关性分析得到 244 个关键词的相似性矩阵。最后将共词矩阵分别导入到 Vosviewer 中进行社会网络分析。为方便对比,输入主题数为 10,即将 244 个关键词分为 10 类,根据 Vosviwer 中关键词 weight 值的不同对每一个类别下的关键词进行类内排序,如表 2 所示:

表 2 基于共词聚类分析的情报学领域不同主题词的优先级排序(部分)

| topic0th:数据分析 | weight 值 | topic1th:网络舆情 | weight 值 | topic2th:文献计量 | weight 值 | topic3th:专利信息 | weight 值 | topic4th:政府开放数据 | weight 值 |
|---|---|---|---|---|---|---|---|---|---|
| 大数据 | 870 | 网络舆情 | 512 | 知识图谱 | 692 | 专利分析 | 338 | 开放数据 | 114 |
| 云计算 | 228 | 微博 | 432 | 文献计量 | 670 | 竞争情报 | 284 | 政府数据 | 94 |
| 情报分析 | 144 | 突发事件 | 268 | 可视化分析 | 658 | 数据挖掘 | 242 | 信息平台 | 64 |
| 信息安全 | 100 | 信息传播 | 146 | 社会网络分析 | 590 | 社会网络 | 196 | 政府数据开放 | 54 |
| 个性化服务 | 76 | 系统动力学 | 132 | 共词分析 | 522 | 因子分析 | 148 | 信息公开 | 48 |
| 智慧城市 | 74 | 复杂网络 | 130 | 聚类分析 | 376 | 文本挖掘 | 124 | 公共服务 | 46 |
| topic5th:数字图书馆服务 | weight 值 | topic6th:电子政务 | weight 值 | topic7th:知识管理 | weight 值 | topic8th:信息服务 | weight 值 | topic9th:科学数据 | weight 值 |
| 数字图书馆 | 330 | 电子政务 | 198 | 知识管理 | 266 | 信息服务 | 294 | 科学数据 | 170 |
| 知识服务 | 310 | 社交媒体 | 132 | 知识共享 | 246 | 微信 | 176 | 机构知识库 | 144 |
| 本体 | 270 | 政府信息公开 | 76 | 社交网络 | 148 | 学科服务 | 166 | 数据管理 | 138 |
| 关联数据 | 186 | 绩效评估 | 72 | 社会化媒体 | 124 | 图书馆服务 | 158 | 数据共享 | 118 |
| 知识发现 | 144 | 信息质量 | 56 | 虚拟社区 | 122 | 移动图书馆 | 150 | 开放获取 | 96 |
| 信息检索 | 124 | 信息推荐 | 56 | 电子商务 | 102 | 信息行为 | 132 | 元数据 | 96 |

为便于表述,将本文提出的基于趋势分析的排序方法称为 A 方法,将基于共词聚类分析的排序方法称为 B 方法。由于 A、B 两种排序方法的理论基础不同,所以两者排序结果存在较大的差异性,表现在:各主题内主题词不同、主题词数量不同、主题标签不同、类内层次不同等。

对比发现,相对 B 方法,A 方法具有以下明显优势:

(1)理论基础的优势。B 方法是基于统计的方法获得学科领域的高频词,忽视了出现在长尾位置的大量低频词和新兴主题词,方法本身带有主观性、不完整性,致使获取的主题不客观、不完整;而 A 方法是利用基于概率推理的 LDA 模型进行主题提取,模型具有严密的数学理论基础,因此,提取的研究主题更全面、更可靠。B 方法虽然综合应用了共词分析、聚类分析与社会网络分析多种方法,但仅是从研究内容的角度考虑了研究者的研究趋势,未考虑读者对文献的需求;而 A 方法通过发文和引文两个维度探析学科主题的研究趋势和需求趋势,进而分析学科主题的研究与利用热度,是文献内容分析与引文分析的有效融合。

(2)聚类层次的优势。B 方法仅是在研究内容单

一层面上将主题进行了聚类;而A方法对研究主题的聚类不仅对学科主题进行了研究内容上的区分,还对各研究主题进行了更细粒度的研究优先级划分,即在研究主题内容划分的基础上,又进一步将每一个研究主题细分为研究贫乏点、研究热点、研究冷点、研究过热点4个等级(部分主题被分为3个等级),是在研究内容和研究等级两个层面上的聚类,聚类效果更加精细。

(3)排序结果的对比。B方法呈现的是对研究主题研究热度的排序,仅能向读者呈现学科领域的研究热点,给出学科领域研究发展的趋势这一种信息,如表2中每个主题下排名靠前的就是该主题的研究热点;而A方法既能展示学科研究热点、研究过热点等,如表1中第②类属于研究热点,第④属于研究过热点;还能给出学科主题的研究等级,如表1中每个主题中 值越高的主题词研究等级越高。可见A方法给出的信息更充分。

(4)聚类性质的优势。B方法属于硬聚类,即一个关键词仅出现在一类主题中;A方法属于软聚类,一个关键词可以出现在不同类中,比如"资源聚合"既属于"topic3th:网络资源"的研究内容,也是"topic5th:开放数据"的研究范畴,该方法与关键词内容指向的多样性是一致的,因此,软聚类的聚类结果更合理。

## 5 研究贡献

本研究的主要贡献在于:

(1)给出相对引文量、发文趋势、引文趋势的定义。相对引文量考虑了发文量对引文量的影响,突破了单纯从引文量看研究主题发展现状的局限性,能够客观地呈现学科领域研究主题的发展趋势。发文趋势反映了学科主题的研究现状,引文趋势反映了研究主题被关注的程度,两者结合,能从研究者和读者两个不同视角分析研究主题的发展态势。

(2)给出研究主题优先级排序方法。本文的排序方法突破了"对所有研究主题进行统一排序"的思路。首先,根据发文趋势和引文趋势分别将不同的研究主题分为4类研究等级,然后,根据给出的排序算法对4类等级下的主题词进行排序。这样不仅能够细致地展示学科领域研究主题的全貌,更能具体呈现学科领域研究主题被研究和被关注的程度。

(3)对我国情报学研究主题进行了研究优先级划分。通过计算主题相似度将情报学研究主题分为10个,利用提出的研究主题优先级排序法将其研究主题划分为贫乏主题、热点主题、冷点主题、过热主题4类等级,并在子主题划分的基础上对10个研究主题进行了研究优先级排序。研究结果可为本学科科研机构制定科研规划、科研人员确定科研方向提供有效、可靠的决策参考。

## 6 讨论

本文是对学科领域研究主题优先级排序的一次尝试性研究,排序算法本身仍有一定的局限性;研究结果的检验问题也没有得到很好的解决,尚需进一步探讨。

(1)在任何一个学科领域中,由于研究者的研究偏好可能发生转移,读者数量也会发生变化,各研究主题的发文量和读者对研究主题的需求量都会发生一定的增减,发文趋势和引文趋势也会随之发生相应的改变;而且,随着学科自身的不断发展和学科交流愈加频繁,还会有新的研究主题不断呈现,上述诸多因素均会影响学科领域研究主题的排序结果。因此,对学科领域研究主题的排序研究应是一个持续性的过程,本文给出的排序方法仅能向相关部门和研究者展示当前的研究态势,仅能为近期的科研选题提供参考。

(2)本文的主题词优先级排序算法,即公式(6)是在本研究数据集上给出的,应用时还需根据具体的数据特点,自行定义。

(3)由于目前没有公认的学科主题排序的验证方法,更未发现有对我国情报学研究主题的排序研究,本研究只是通过对同一组数据集进行共词聚类分析,在理论基础、聚类性质、聚类层次、排序结果等方面与本文学科主题优先级排序法进行了比较,在对比分析中突显本文排序方法的优势,但排序结果与专家的判断是否一致本文并未给予合理的验证。

**参考文献:**

[1] 张晓林. 颠覆数字图书馆的大趋势[J]. 中国图书馆学报,2011,37(195):4-11.

[2] 刘欣. 基于阅读价值的科技文献排序方法研究[D]. 大连:大连理工大学, 2010.

[3] 王燕鹏. 国内基于主题模型的科技文献主题发现及演化研究进展[J]. 图书情报工作, 2016, 60(3):130-137.

[4] 关鹏,王曰芬. 基于LDA主题模型和生命周期理论的科学文献主题挖掘[J]. 情报学报,2015, 34(3):286-299.

[5] DONOHUE J C. Understanding scientific literature: a bibliographic approach[M]. Massachusetts: The MIT Press,1973.

[6] 唐果媛,张薇. 基于共词分析法的学科主题演化研究进展与分析[J]. 图书情报工作,2015,59(5):128-136.

[7] 陈仕吉,王小梅. 基于C-value与TF-IDF的文献簇主题识别研究[J]. 情报学报, 2009,28(6):821-826.

[8] 陈勇跃,田文芳,吴金红. 主题领域研究热点跟踪及趋势预测的可

视化分析方法研究[J].情报理论与实践,2017,40(6):117－121.
[9] 巴志超,李纲,朱世伟.共现分析中的关键词选择与语义度量方法研究[J].情报学报,2016,35(2):197－207.
[10] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11):613－620.
[11] DEERWESTER S C. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science,1990,41(6):391－407.
[12] Probabilistic latent semantic analysis[EB/OL].[2019－02－26]. https://en.wikipedia.org/wiki/Latent_semantic_analysis#Latent_semantic_indexing.
[13] Latent dirichlet allocation[EB/OL].[2019－02－26]. https://max.book118.com/html/2017/0724/123956096.shtm.
[14] Topics over time: a non-Markov continuous-time model of topical trends[EB/OL].[2019－02－26]. https://wenku.baidu.com/view/a1b8c42d4b73f242336c5fce.html.
[15] Extracting multilingual topics from unaligned comparable corpora [EB/OL].[2019－02－26]. https://link.springer.com/chapter/10.1007%2F978-3-642-12275-0_39.
[16] 李湘东,巴志超,黄莉.一种基于加权LDA模型和多粒度的文本特征选择方法[J].现代图书情报技术,2015,31(5):42－48.
[17] 王曰芬,傅柱,陈必坤.采用LDA主题模型的国内知识流研究结构探讨:以学科分类主题抽取为视角[J]. 现代图书情报技术,2016,32(4):8－19.
[18] 王连喜,曹树金.学科交叉视角下的网络舆情研究主题比较分析——以国内图书情报学和新闻传播学为例[J].情报学报,2017,36(2):159－169.
[19] 许腾腾,黄恒君.一种改进的Supervised-LDA文本模型及其应用[J].计算机工程,2018,44(1): 69－73,78.
[20] 曲靖野,陈震,胡铁楠.共词分析与LDA模型分析在文本主题挖掘中的比较研究[J].情报科学,2018,36(2):18－23.
[21] WANG C H, ZHANG M, RU L Y, et al. Automatic online news topic ranking using media focus and user attention based on aging theory[C]// Proceedings of the 17th ACM conference on information and knowledge management. New York: ACM Press, 2008: 1033－1042.
[22] 姜晓伟,王建民,丁贵广.基于主题模型的微博重要话题发现与排序方法[J].计算机研究与发展,2013,50(5):179－185.
[23] 刘培玉,侯秀艳,朱振方.基于热度联合排序的微博热点话题发现[J].计算机科学与探索,2016, 10(4):574－581.
[24] CUI W, LIU S, TAN L, et al. Text flow: towards better understanding of evolving topics in text[J]. IEEE transactions on visualization & computer graphics,2011,17(12):2412－2421.
[25] 肖智博.排序主题模型及其应用研究[D].大连:大连海事大学,2014.
[26] XIAO Z, CHE F, MIAO E, et al. Increasing serendipity of recommender system with ranking topic model[J]. Applied mathematics & information sciences, 2014,8(4):2041－2053.
[27] 蒋卓人,高良才,赵星.中英文科技主题排序相关性的比较研究:以计算机领域为例[J]. 情报学报,2017, 36(9):940－953.
[28] ALAM M M, ISMAIL M A. RTRS: a recommender system for academic researchers[J]. Scientometrics, 2017,113(1):1－24.

**作者贡献说明:**

李秀霞:提出研究思路,设计研究方案,撰写论文;
程结晶:设计研究方案与修改论文;
韩霞:收集与处理数据。

## The Prioritization of Subject Research Topics Based on the Integration of Writing Trends and Citation Trends: Taking the Subject of Information Science in China as an Example

Li Xiuxia[1] Cheng Jiejing[2] Han Xia[1]
[1] School of Communication, Qufu Normal University, Rizhao 276826
[2] Institute of Social Development, Yangzhou University, Yangzhou 225008

**Abstract:** [**Purpose/significance**] Topic sorting is not only the basic problem for information retrieval and information organization, but also an important work of subject service. The effective sorting of subject field research topics can help researchers and decision-making departments to grasp the research situation of the subject field effectively, locate the direction of scientific research accurately and make scientific research decisions quickly. [**Method/process**] This paper proposes the prioritization algorithm based on the combination of topic extraction and trend analysis. Then it takes the research topics of Library and Information Science as an example to extract the research topics of the sample literature, and each research topic is divided into four sub-topics: poor theme, hot topic, cold point theme, and overheated topic. Next priority ranking is carried out in subclasses. [**Result/conclusion**] The empirical results show that the priority ranking algorithm can display the development level of research topics in an all-round, fine-grained and deep way. This method provides a new perspective for realizing dynamic intelligence analysis from time dimension.

**Keywords:** writing trend citation trend research topic prioritization